

# **COMPARING THE PERFORMANCE OF MPI ON THE CRAY T3E-900, THE CRAY ORIGIN 2000 AND THE IBM P2SC**

Glenn R. Luecke and James J. Coyle  
grl@iastate.edu and jjc@iastate.edu

Iowa State University  
Ames, Iowa 50011-2251, USA

October 28, 1997

## **Abstract**

This study was conducted to evaluate relative communication performance of the Cray T3E-900, the Cray Origin 2000 and the IBM P2SC on a collection of 13 communication tests. These tests call MPI routines using 2 to 64 processors with messages varying from 8 Bytes to 10 MB. The relative performance of these machines varied depending on the communication test, but overall the T3E-900 was often 2 to 4 times faster than the Origin 2000 and P2SC. The Origin 2000 and P2SC performed about the same for most of the tests.

## INTRODUCTION

The performance of the communication network of a parallel computer plays a critical role in its overall performance, see [4,6]. Writing scientific programs with calls to MPI [5,9] routines is rapidly becoming the standard for writing programs with explicit message passing. Thus, to evaluate the performance of the communication network of a parallel computer for scientific computing, a collection of communication tests that use MPI for the message passing have been written. These communication tests are a significant enhancement from those used in [7] and have been designed to test those communication patterns that we feel are likely to occur in scientific programs. This paper reports the results of these tests on the Cray T3E-900, the Cray Origin 2000 and the IBM P2SC ("Power 2 Super Chip").

## DESCRIPTION OF THE COMMUNICATION TESTS AND RESULTS

All tests were written in Fortran with calls to MPI routines for the message passing. The communication tests were run with message sizes ranging from 8 bytes to 10 MB and with the number of processors ranging from 2 to 64. Because of memory limitations, for some of the tests the 10 MB message size was replaced by a message of size 1 MB. Some of these communication patterns took a very short amount of time to execute so they were looped to obtain a wall-clock time of at least one-second in order to obtain more accurate timings. The time to execute a particular communication pattern was then obtained by dividing the total time by the number of loops. All timings were done using the MPI wall-clock timer, `mpi_wtime()`. A call to `mpi_barrier` was made just prior to the first call to `mpi_wtime` and again just prior to the second call to `mpi_wtime` to ensure processor synchronization for timing. Sixty-four bit real precision was used on all machines. Tests run on the Cray T3E-900 and Cray Origin 2000 were executed on machines dedicated to running only our tests. For these machines, five runs were made and the best performance results are reported. Tests run on the IBM P2SC were executed using LoadLeveler so that only one job at a time would be executing on the 32 nodes used. However, jobs running on other nodes would sometimes cause variability in the data so the tests were run at least ten times and the best performance numbers reported.

Tests for the **Cray T3E-900** were run on a 64-processor machine located in Chippewa Falls, Wisconsin. Each processor is a DEC EV5 microprocessor running at 450 MHz with peak theoretical performance of 900 Mflop/s. The three-dimensional, bi-directional torus communication network of the T3E-900 has a bandwidth of 350 MB/second and latency of 1.5 microsecond. For more information on the T3E-900 see [9]. The UNICOS/mk version 1.3.1 operating system, the cf90 version 3.0 Fortran compiler with the `-O2 -dp` compiler options and

MPI version 3.0 were used for these tests. The MPI implementation used had the hardware data streams work-around enabled even though this is not needed for the T3E-900.

Tests for the **Cray Origin 2000** were run on a 128-processor machine located in Eagan, Minnesota. Each processor is a MIPS R10000, 195 MHz microprocessor with a peak theoretical performance of 390 Mflop/s. Each node consists of two processors sharing a common memory. The communication network is a hypercube for up to 32 processors and is called a “fat bristled hypercube” for more than 32 processors since multiple hypercubes are interconnected via the CrayRouter. There is one port from the memory in a node that is shared by the two node processors and this port has a bandwidth of 780 MB per second and a latency of about 300 nanoseconds. Each node has an interface for an incoming stream of data and an outgoing stream of data that can operate concurrently for passing data from one node to another node. Each of these streams has a peak bandwidth of 780 MB per second making the peak node-to-node bandwidth 1.56 GB per second. The maximum remote latency in a 128-processor system is about 1 microsecond. For more information see [9]. A pre-release version of the Irix 6.5 operating system, MPI from version 1.1 of the Message Passing Toolkit, and the MipsPro 7.20 Fortran compiler with -O2 -64 compiler options were used for these tests.

Tests for the **IBM P2SC** were run at the Maui High Performance Computing Center. The peak theoretical performance of each of these processors is 480 Mflop/s for the 120 MHz thin nodes and 540 Mflop/s for the 135 MHz wide nodes. The communication network has a peak bi-directional bandwidth of 150 MB/second with a latency of 40.0 microseconds for thin nodes and 39.2 microseconds for wide nodes. Performance tests were run on thin nodes each with 128 MB of memory. At the Maui High Performance Computing Center, there were only 48 thin nodes available for running these tests, so there is no data for 64 processors. For more information about the P2SC see [10]. The AIX 4.1.5.0 operating system, xlf version 4.1.0.0 Fortran compiler with -O3 qarch=pwr2 compiler options, and MPI version 2.2.0.2 were used for these tests.

### **Communication Test 1** (point-to-point, see table 1)

The first communication test measures the time required to send a real array  $A(1:n)$  from one processor to another by dividing by two the time required to send  $A$  from one processor to another and then send  $A$  back to the original processor, where  $n$  is chosen to obtain a message of the desired size. Thus, to obtain a message of size 1 KB,  $n = 1,000/8 = 125$ . Since each  $A(i)$  is 8 Bytes, the communication rate for sending a message from one processor to another is calculated by  $2 \cdot 8 \cdot n / (\text{wall-clock time})$ , where the wall-clock time is the time to send  $A$  from one processor to another and then back to the original processor and where  $n$  is chosen to obtain the desired message size. This test is the same as the COMMS1 test described in section 3.3.1 of [4]. This test uses `mpi_send` and `mpi_recv`. Table 1 gives performance rates in KB per second. As is done in all the tables, the last column gives the ratios of the performance results of the T3E-900 to the IBM P2SC and of the T3E-900 to the Origin 2000. Notice that the achieved bandwidth on this

test is significantly less than the bandwidth rates provided by the vendors: 350,000 KB for the T3E-900, 150,000 KB for the P2SC, and 750,000 KB for the Origin.

Message Size (Bytes)	T3E-900	IBM P2SC	Origin 2000	T3E/IBM,T3E/Origin
8	371	215	303	1.7, 1.2
1,000	30087	14475	18580	2.1, 1.6
100,000	144337	78726	86475	1.8, 1.7
10,000,000	151113	103162	90880	1.5, 1.7
Peak Rates	350000	150000	750000	

**Table 1:** Point-to-point communication rates plus advertised Peak Rates in KB/second.

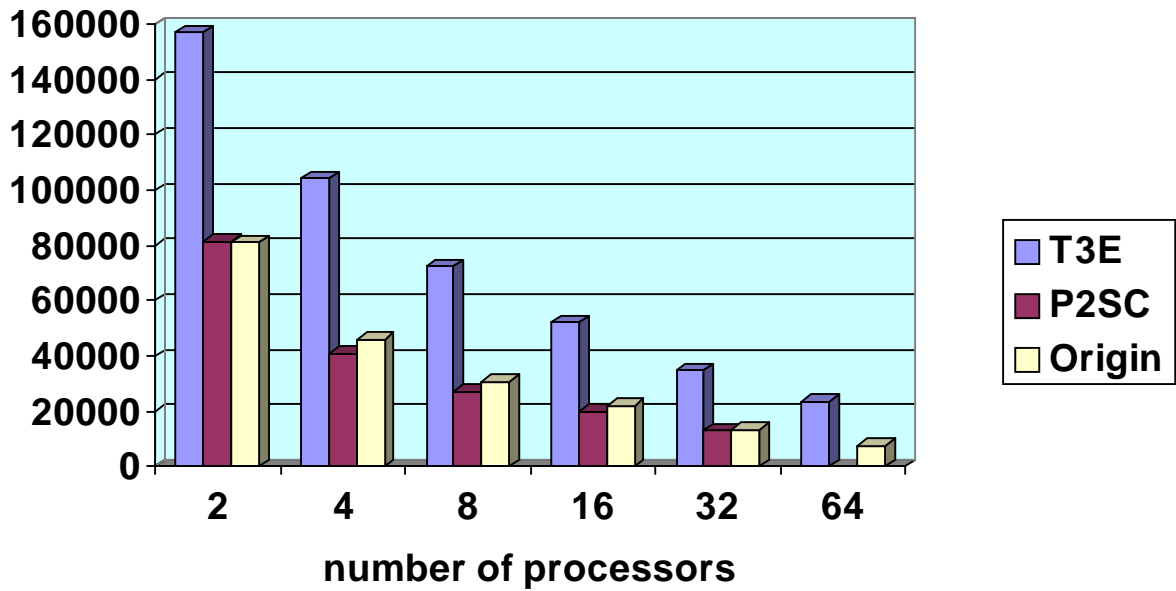
**Communication Test 2** (broadcast, see tables 2-a, 2-b and 2-c and figures 1 and 2):

This test measures communication rates for sending a message from one processor to all other processors and uses the `mpi_bcast` routine. This test is the COMMS3 test described in [4]. To better evaluate the performance of this broadcast operation, define a **normalized broadcast rate** as

$$(\text{total data rate})/(p-1)$$

where  $p$  is the number of processors involved in the communication and **total data rate** is the total amount of data sent on the communication network per unit time measured in KB per second. Let  $R$  be the data rate when sending a message from one processor to another and let  $D$  be the total data rate for broadcasting the same message to the  $p-1$  other processors. If the broadcast operation and communication network were able to concurrently transmit the messages, then  $D = R*(p-1)$  and thus the normalized broadcast rate would remain constant as  $p$  varied for a given message size. Therefore, for a fixed message size, the rate at which the normalized broadcast rate decreases as  $p$  increases indicates how far the broadcast operation is from being ideal. Assume the real array  $A(1:n)$  is broadcast from the root processor where each  $A(i)$  is 8 Bytes, then the communication rate is calculated by  $8*n*(p-1)/(\text{wall-clock time})$  and then normalized by dividing by  $p-1$  to obtain the normalized broadcast rate.

Table 2-a gives the normalized broadcast rates obtained by keeping the root processor fixed for all repetitions of the broadcast. Figure 1 shows the graph of these results for a message of size 100 KB. Notice that for all machines for a fixed message size the normalized broadcast rate decreases as the number of processors increase (instead of being constant). Notice that on this test the P2SC and Origin machines perform roughly about the same and that the T3E-900 ranges from 1.1 to 3.0 times faster than the P2SC and Origin. Observe that the Origin does not scale well as the number of processors increase as compared with the T3E-900. One might expect that the communication rate for a broadcast with 2 processors would be the same as the rate for communication test 1. However, the rates measured for the broadcast in tables 2-a and 2-b are higher than those measured in communication test 1 for all machines. It is not clear why this is so.



**Figure 1:** Normalized broadcast rates for a 100 KB message (from table 2-a).

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	545	350	518	1.6, 1.1
8	4	388	194	245	2.0, 1.6
8	8	290	132	159	2.2, 1.8
8	16	264	103	110	2.6, 2.4
8	32	223	83	60	2.7, 3.7
8	64	196	NA	42	---, 4.7
1,000	2	43613	32795	26703	1.3, 1.6
1,000	4	24439	16206	13913	1.5, 1.8
1,000	8	19070	11092	9504	1.7, 2.0
1,000	16	15898	8702	6647	1.8, 2.4
1,000	32	13338	7089	4958	1.9, 2.7
1,000	64	11764	NA	3819	---, 3.1
100,000	2	156974	81597	80976	1.9, 1.9
100,000	4	104085	41035	46231	2.5, 2.3
100,000	8	72496	27302	30935	2.7, 2.3
100,000	16	52277	20138	22145	2.6, 2.4
100,000	32	34894	13145	13340	2.7, 2.6
100,000	64	23805	NA	7904	---, 3.0
10,000,000	2	156974	103183	92767	1.5, 1.7
10,000,000	4	80423	51448	45116	1.6, 1.8
10,000,000	8	50799	34161	28853	1.5, 1.8
10,000,000	16	37636	25636	19522	1.5, 1.9
10,000,000	32	29212	15359	13334	1.9, 2.2
10,000,000	64	24511	NA	8078	---, 3.0

**Table 2-a:** Normalized broadcast rates in KB/second with a fixed root processor.

Table 2-b gives the normalized broadcast rates where the root processor is cycled through all p processors as the broadcast operation is repeated. Notice that the rates do change from those in table 2-a, but the maximum percent change depends on the machine. The maximum percent change is about 14% for the T3E-900, 150% for the P2SC, and about 50% for the Origin.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	637	178	415	3.6, 1.5
8	4	387	126	213	3.1, 1.8
8	8	287	93	178	3.1, 1.6
8	16	247	77	113	3.2, 2.2
8	32	211	62	58	3.4, 3.6
8	64	185	NA	37	---, 5.0
1,000	2	40271	13035	20186	3.1, 2.0
1,000	4	27030	9665	11965	2.8, 2.3
1,000	8	18421	7503	9335	2.5, 2.0
1,000	16	15273	6218	6711	2.6, 2.3
1,000	32	12866	4943	4920	2.6, 2.6
1,000	64	11419	NA	3566	---, 3.2
100,000	2	159953	77769	86966	2.1, 1.8
100,000	4	95802	40145	40757	2.4, 2.4
100,000	8	70000	26871	28524	2.6, 2.5
100,000	16	48677	19968	20357	2.4, 2.4
100,000	32	33877	12719	11779	2.7, 2.9
100,000	64	23671	NA	6022	---, 3.9
10,000,000	2	162152	103174	87630	1.6, 1.9
10,000,000	4	73311	51422	45019	1.4, 1.6
10,000,000	8	50964	34284	28741	1.5, 1.8
10,000,000	16	36312	25684	19531	1.4, 1.9
10,000,000	32	29265	15233	12219	1.9, 2.4
10,000,000	64	24575	NA	8269	---, 3.0

**Table 2-b:** Normalized broadcast rates in KB/second with the root processor cycled.

To better understand the amount of concurrency occurring in the broadcast operation, define the **log normalized broadcast rate** as

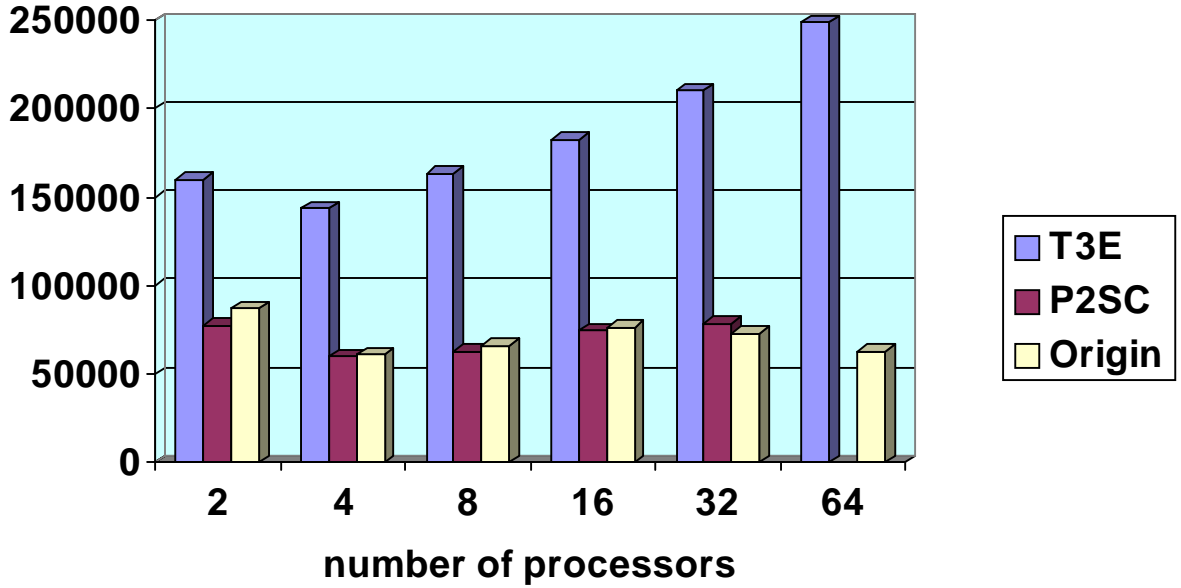
$$(\text{total data rate})/\log(p)$$

where p is the number of processors involved in the communication and  $\log(p)$  is the log base 2 of p. Thus, if binary tree parallelism were being utilized, the log normalized data rate would be constant for a given message size as p varies. Table 2-c gives the log normalized data rates with

a fixed root processor and shows in fact that concurrency is being utilized in the broadcast operation for these machines. Figure 2 shows these results for a message of size 100 KB. Notice that the performance of the T3E-900 is significantly better than binary tree parallelism for all message sizes tested. For messages of size 8 Bytes and 1 KB, the P2SC performs better than binary tree parallelism and yields binary tree parallelism for the other two message sizes. The Origin gives better than binary tree parallelism for a 1 KB message and binary tree parallelism for the other three message sizes.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM,T3E/Origin
8	2	637	178	415	3.6, 1.5
8	4	580	189	319	3.1, 1.8
8	8	669	217	415	3.1, 1.6
8	16	926	288	423	3.2, 2.2
8	32	1308	384	353	3.4, 3.7
8	64	1942	NA	388	---, 5.0
1,000	2	40271	13035	20186	3.1, 2.0
1,000	4	40545	14497	17947	2.8, 2.3
1,000	8	42982	17507	21781	2.5, 2.0
1,000	16	57273	23317	25166	2.5, 2.3
1,000	32	79769	30646	30504	2.6, 2.6
1,000	64	119899	NA	37443	---, 3.2
100,000	2	159953	77769	86966	2.1, 1.8
100,000	4	143703	60217	61135	2.4, 2.4
100,000	8	163333	62699	66556	2.6, 2.5
100,000	16	182538	74880	76338	2.4, 2.4
100,000	32	210037	78857	73029	2.7, 2.9
100,000	64	248545	NA	63231	---, 3.9
10,000,000	2	162152	103174	87630	1.6, 1.9
10,000,000	4	109966	77133	67528	1.4, 1.6
10,000,000	8	118916	79996	67062	1.5, 1.8
10,000,000	16	136170	96315	73241	1.4, 1.9
10,000,000	32	183675	94444	75757	1.9, 2.4
10,000,000	64	258037	NA	86824	---, 3.0

**Table 2-c:** Log normalized broadcast rates in KB/second with the root processor cycled.



**Figure 2:** Log normalized broadcast rates for a 100 KB message (from table 2-c).

### Communication Test 3 (reduce, see table 3):

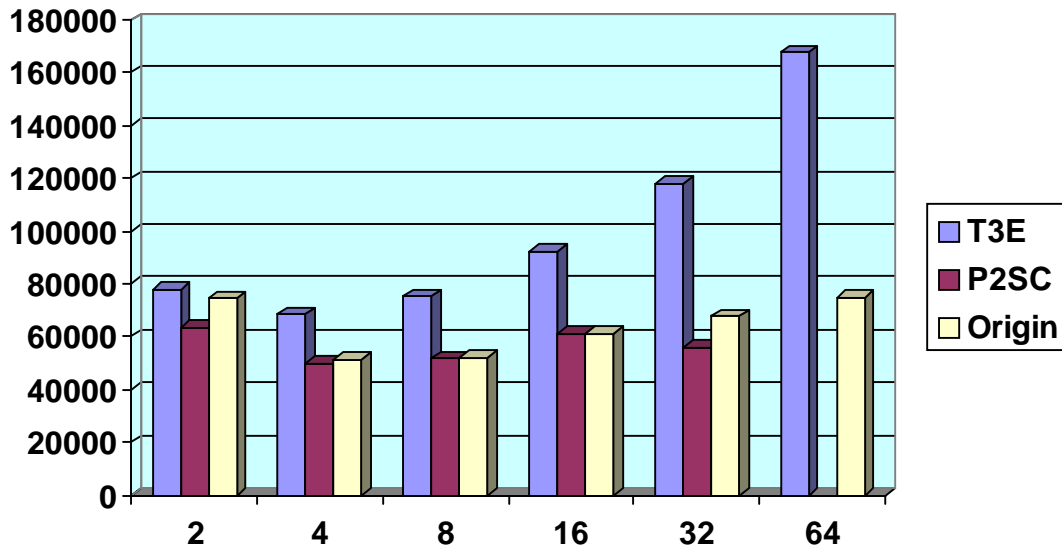
Assume that there are  $p$  processors and that processor  $i$  has a message,  $A_i(1:n)$ , for  $i = 0, p-1$  and where  $n$  is chosen to obtain the message of the desired size. Test 3 measures communication rates for varying sizes of the  $A_i$ 's when calculating  $A = \sum A_i$  and placing  $A$  on the root processor. Thus, this test uses `mpi_reduce` with the `mpi_sum` option. Since each element of  $A_i$  is 8 bytes, the communication rate can be calculated by  $8*n*(p-1)/(\text{wall-clock time})$  and then normalized by dividing by  $p-1$ . As was done with `mpi_bcast`, one could also calculate a log normalized data rate. Table 3 contains log normalized data rates since, as was true for `mpi_bcast`, more information is obtained. Table 3 shows that the Origin exhibits binary tree parallelism and the other two machines exhibit better than binary tree parallelism. Notice that the T3E-900 performs well compared with the two other machines for messages of size 8 Bytes and 1 KB. However, for messages of size 100 KB (with 8 or more processors) and 10 MB (with 4 or more processors), the IBM machine gives superior performance. This may be due to the optimization method used by IBM for these larger messages in their implementation of `mpi_reduce`.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	439	187	430	2.3, 1.0
8	4	336	185	372	1.8, .90
8	8	385	145	345	2.7, 1.1
8	16	491	120	368	4.1, 1.3
8	32	670	74	360	9.0, 1.9
8	64	977	NA	431	---, 2.2
1,000	2	29048	19375	26240	1.5, 1.1
1,000	4	22568	14783	21926	1.5, 1.0
1,000	8	24981	11909	21147	2.1, 1.2
1,000	16	30848	15799	25095	2.0, 1.2
1,000	32	42259	20386	28991	2.1, 1.5
1,000	64	59220	NA	38294	---, 1.5
100,000	2	48988	38323	63074	1.3, .78
100,000	4	40827	39591	47393	1.0, .86
100,000	8	43479	52603	47096	.83, .92
100,000	16	53486	75945	53141	.70, 1.0
100,000	32	70599	88666	59477	.80, 1.2
100,000	64	96264	NA	67862	---, 1.2
10,000,000	2	49599	39405	41888	1.3, 1.2
10,000,000	4	40139	56601	29085	.71, 1.4
10,000,000	8	41720	78647	26752	.53, 1.6
10,000,000	16	51990	118088	28898	.44, 1.8
10,000,000	32	70519	162291	35104	.43, 2.0
10,000,000	64	94133	NA	32414	---, 2.9

**Table 3:** Log normalized data rates in KB/second for mpi\_reduce with the mpi\_sum option.

**Communication Test 4** (all reduce, see table 4 and figure 3):

This communication test is the same as communication test 3 except **A** is placed on all processors instead of only on the root processor. This test uses the `mpi_allreduce` routine and is functionally equivalent to a reduce followed by a broadcast. Thus, the communication rate for this test is calculated by  $2*[8*n*(p-1)]/(\text{wall-clock time})$  and then divided by  $p-1$  to get a normalized data rate. Since normalized data rates drop sharply for fixed message sizes as the number of processors increase, more information is obtained by calculating log normalized data rates, see table 4 and figure 3. Notice that the P2SC and Origin exhibit binary tree parallelism and the T3E does much better. Also notice that for most of the cases in test 4, the T3E-900 significantly outperforms the other two machines. The P2SC does not scale nearly as well as the T3E-900 for messages of sizes 8 Bytes and 1 KB. The Origin does not scale nearly as well as the T3E-900 for all message sizes.



**Figure 3:** Log normalized data rates for `mpi_allreduce` for a 1 KB message (from table 4).

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM,T3E/Origin
8	2	671	225	363	3.0, 1.8
8	4	671	191	291	3.5, 2.3
8	8	775	198	299	3.9, 2.6
8	16	994	233	319	4.3, 3.1
8	32	1364	273	291	5.0, 4.7
8	64	2006	NA	357	---, 5.6
1,000	2	45958	17741	20312	2.6, 2.3
1,000	4	39863	13979	13749	2.9, 2.9
1,000	8	44707	14191	14366	3.2, 3.1
1,000	16	55024	16695	17561	3.3, 3.1
1,000	32	73352	18953	20175	3.9, 3.6
1,000	64	105851	NA	26345	---, 4.0
100,000	2	77926	63422	74396	1.2, 1.0
100,000	4	68282	49868	51500	1.4, 1.3
100,000	8	75082	51732	52309	1.5, 1.4
100,000	16	92344	61354	61005	1.5, 1.5
100,000	32	118023	56104	67729	2.1, 1.7
100,000	64	167832	NA	74676	---, 2.2
10,000,000	2	76511	69839	47992	1.1, 1.6
10,000,000	4	67659	55781	28139	1.2, 2.4
10,000,000	8	73920	59689	27445	1.2, 2.7
10,000,000	16	92348	71348	32753	1.3, 2.8
10,000,000	32	114210	70184	38762	1.6, 2.9
10,000,000	64	166436	NA	33275	---, 5.0

**Table 4:** Log normalized data rates in KB/second for mpi\_allreduce with the mpi\_sum option.

**Communication Test 5** (gather, see table 5):

Assume that there are  $p$  processors and that processor  $i$  has a message,  $A_i(1:n)$ , for  $i = 0, p-1$ . This test uses the mpi\_gather routine and measures the communication rate for gathering the  $A_i$ 's into an array  $B$  located on the root processor, where  $B(1:n,i) = A_i(1:n)$  for  $i = 0, p-1$ . Since the normalized data rates drop sharply as the number of processors increase for a fixed message size, the log normalized data rate provides more information and is used for reporting performance results for this test. Thus the communication rate is calculated by  $8*n*(p-1)/(wall-clock-time)$  and then normalized by dividing by  $\log(p)$ . Because of the large amount of memory

required to store **B** when a large number of processors is used, the largest message size used for this test was 1 MB instead of 10 MB. Relative performance results are quite mixed but the T3E-900 outperformed the other machines on all of these tests. Notice the large drop in performance on the Origin for 8 Byte and 1 KB messages as the number of processors increase.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	191	175	222	1.1, 0.9
8	4	266	167	233	1.6, 1.1
8	8	268	107	163	2.5, 1.6
8	16	251	83	83	3.0, 3.0
8	32	223	62	37	3.6, 6.0
8	64	189	NA	11	---, 18.
1,000	2	16195	14253	13668	1.1, 1.2
1,000	4	22745	8988	12096	2.5, 1.9
1,000	8	19873	3607	10376	5.5, 1.9
1,000	16	18930	4223	7189	4.5, 2.6
1,000	32	14706	5090	4278	2.9, 3.4
1,000	64	12348	NA	1449	---, 8.5
100,000	2	43333	30717	41083	1.4, 1.1
100,000	4	48399	34395	32727	1.4, 1.5
100,000	8	31943	21560	27804	1.5, 1.5
100,000	16	34436	23603	22256	1.5, 1.5
100,000	32	28632	19604	16343	1.5, 1.8
100,000	64	26418	NA	10616	---, 2.5
1,000,000	2	43587	37078	39117	1.2, 1.1
1,000,000	4	49061	38486	28221	1.3, 1.7
1,000,000	8	42240	32907	20090	1.3, 2.1
1,000,000	16	35696	27101	16061	1.3, 2.2
1,000,000	32	24118	23305	11960	1.0, 2.0
1,000,000	64	NA	NA	8442	

**Table 5:** Log normalized data rates in KB/second for mpi\_gather.

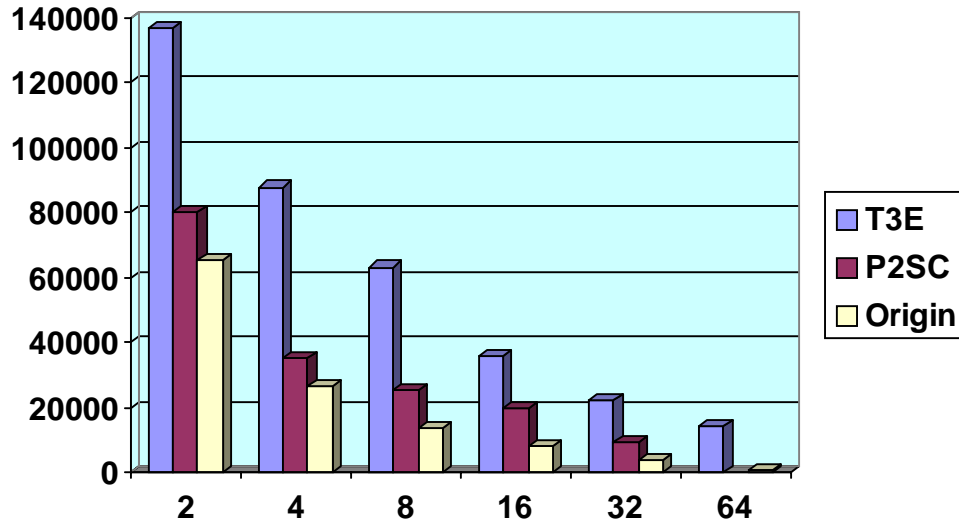
**Communication Test 6** (all gather, see table 6 and figure 4):

This test is the same as test 5 except the gathered message is placed on all processors instead of only on the root processor. Test 6 is functionally equivalent to a gather followed by a

broadcast and uses `mpi_allgather`. The communication rate calculated by  $2*[8*n*(p-1)]/(\text{wall-clock time})$  and is divided by  $\log(p)$  to obtain a log normalized data rate. Because of the large amount of memory required to store **B** when a large number of processors is used, the largest message size used for this test was 1 MB. Notice the large drop in relative performance of the Origin as the number of processors increase. Also notice that none of these machines were able to achieve binary tree parallelism on this test.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	658	270	426	2.4, 1.5
8	4	498	203	323	2.5, 1.5
8	8	397	208	252	1.9, 1.6
8	16	315	244	176	1.3, 1.8
8	32	260	273	112	1.0, 2.3
8	64	221	NA	32	---, 7.0
1,000	2	50619	20732	20673	2.4, 2.4
1,000	4	36569	13133	14007	2.8, 2.6
1,000	8	27246	8815	9837	3.1, 2.8
1,000	16	18555	7729	5726	2.4, 3.2
1,000	32	12189	5723	3813	2.1, 3.2
1,000	64	9251	NA	2594	---, 3.6
100,000	2	136782	80191	65571	1.7, 2.1
100,000	4	87845	35115	26501	2.5, 3.3
100,000	8	63273	25580	14028	2.5, 4.5
100,000	16	36139	19988	8310	1.8, 4.3
100,000	32	22184	9226	3776	2.4, 5.9
100,000	64	14238	NA	1019	---, 14.
1,000,000	2	137141	93250	60310	1.5, 2.3
1,000,000	4	91551	38666	20747	2.4, 4.4
1,000,000	8	69078	26796	9011	2.6, 7.7
1,000,000	16	36694	19778	3878	1.9, 9.5
1,000,000	32	21117	9393	2027	2.2, 10.
1,000,000	64	NA	NA	725	

**Table 6:** Log normalized data rates in KB/second for `mpi_allgather`.



**Figure 4:** Log normalized data rates for `mpi_allgather` for a 100 KB message (from table 6).

**Communication Test 7** (scatter, see table 7):

Assume that  $\mathbf{B}$  is a two dimensional array,  $\mathbf{B}(1:n,0:p-1)$ , where  $p$  is the number of processors used. This test uses `mpi_scatter` and measures communication rates for scattering  $\mathbf{B}$  from the root processor to all other processors so that processor  $j$  receives  $\mathbf{B}(1:n,j)$ , for  $j = 0, p-1$ . The communication rate for this test is calculated by  $8 \cdot n \cdot (p-1) / (\text{wall-clock-time})$  and then dividing by  $\log(p)$  to obtain the log normalized data rate. Because of the large memory requirements when a large number of processors is used for this test, the largest message used for this test was 1 MB.

Notice that relative to the T3E-900, the Origin performance results decrease as the number of processors increase for each message size. This also happens for the P2SC for all message sizes other than 8 Bytes. Observe that the Origin and IBM P2SC perform roughly the same for most cases and that the T3E-900 is 2 to 3 times faster than both of these machines for most tests. Also notice that none of these machines are able to achieve binary tree parallelism except for the P2SC on the 8 Byte message.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	417	314	265	1.3, 1.6
8	4	423	293	224	1.4, 1.9
8	8	418	345	196	1.2, 2.1
8	16	356	510	165	0.7, 2.2
8	32	304	614	136	0.5, 2.2
8	64	263	NA	105	---, 2.5
1,000	2	33628	28164	19732	1.2, 1.7
1,000	4	30483	20105	12077	1.5, 2.5
1,000	8	25452	11104	9513	2.3, 2.7
1,000	16	21326	9638	7613	2.2, 2.8
1,000	32	17992	7682	6026	2.3, 3.0
1,000	64	15687	NA	4599	---, 3.4
100,000	2	91932	65445	67439	1.4, 1.4
100,000	4	83205	43809	46038	1.9, 1.8
100,000	8	77709	30333	33124	2.6, 2.3
100,000	16	66645	22748	26280	2.9, 2.5
100,000	32	55775	17943	20975	3.1, 2.7
100,000	64	46074	NA	16023	---, 2.9
1,000,000	2	93528	73781	66539	1.3, 1.4
1,000,000	4	84501	46050	43124	1.8, 2.0
1,000,000	8	78738	32604	30854	2.4, 2.6
1,000,000	16	67384	25155	23554	2.7, 2.9
1,000,000	32	56873	20367	19425	2.8, 2.9
1,000,000	64	46883	NA	15330	---, 3.1

**Table 7:** Log normalized data rates in KB/second for mpi\_scatter.

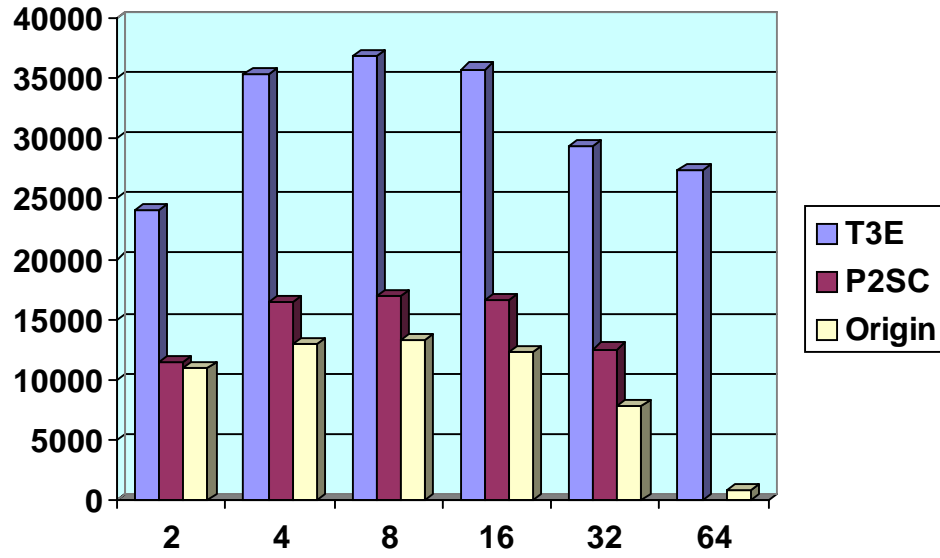
**Communication Test 8** (all-to-all, see table 8, figure 5):

Assume **C** is a three dimensional array, **C**(1:n,0:p-1,0:p-1) with **C**(1:n,j,0:p-1) on processor j. Also assume that **C**(1:n,j,k) is sent to processor k, where j and k both range from 0 to p-1. This test uses mpi\_alltoall and the communication rate is calculated by  $8 \cdot n \cdot (p-1) \cdot p / (\text{wall-clock time})$  and then normalized by dividing by p and not by log(p). As the number of processors increase, this test provides a good stress test for the communication network. Because of the large memory requirements when a large number of processors are used for this test, the largest message used for this test was 1 MB.

Notice that table 8 and figure 5 use normalized data rates and not log normalized data rates. Thus, table 8 and figure 5 show the high level of parallelism achieved for mpi\_alltoall for these machines, especially for the T3E-900. Also notice that relative to one another, the performance of the T3E-900 and P2SC remained nearly constant for all these tests with the T3E-900 giving roughly twice the performance of the P2SC. However, the performance of the Origin relative to the other two machines dropped significantly as the number of processors increases. There was insufficient memory on the T3E-900 to run this test for a 1 MB message with 64 processors.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	305	151	154	2.0, 2.0
8	4	480	186	210	2.6, 2.3
8	8	574	273	203	2.1, 2.8
8	16	615	405	135	1.5, 3.0
8	32	620	527	62	1.2, 10.
8	64	630	NA	25	---, 25.
1,000	2	24129	11402	10880	2.1, 2.2
1,000	4	35355	16473	12927	2.1, 2.7
1,000	8	36890	16926	13251	2.2, 2.8
1,000	16	35775	16590	12315	2.2, 2.9
1,000	32	29419	12524	7843	2.3, 3.8
1,000	64	27342	NA	882	---, 31.
100,000	2	63885	40787	34971	1.6, 1.8
100,000	4	75318	44502	42933	1.7, 1.8
100,000	8	88095	44870	38038	2.0, 2.3
100,000	16	70380	44775	27105	1.6, 2.6
100,000	32	52421	33790	14384	1.6, 3.6
100,000	64	42651	NA	315	---, 135
1,000,000	2	61966	46458	38595	1.3, 1.6
1,000,000	4	92355	51498	36999	1.8, 2.5
1,000,000	8	103173	52906	29519	2.0, 3.5
1,000,000	16	73065	52725	8655	1.4, 8.4
1,000,000	32	53258	43369	2635	1.2, 20.
1,000,000	64	NA	NA	945	

Table 8: Normalized data rates in KB/second for mpi\_alltoall.



**Figure 5:** Normalized data rates for mpi\_alltoall for a 1 KB message (from table 8).

**Communication Test 9** (broadcast-gather, see table 9):

This test uses `mpi_bcast` and `mpi_gather` and measures communication rates for broadcasting a message from the root processor to all other processors and then having the root processor gather these messages back from all processors. This test is included since there may be situations where the root processor will broadcast a message to the other processors, the other processors use this message to perform some calculations, and then the newly computed data is gathered back to the root processor. The communication rate is calculated by  $2 \cdot [8 \cdot n \cdot (p-1)] / (\text{wall-clock time})$  and then divided by  $\log(p)$  to obtain the log normalized data rate. Because of the large memory requirements when a large number of processors is used, the largest message used for this test was 1 MB.

Notice that the T3E-900 significantly outperforms the other machines. Also observe that there seems to be a problem on the Origin for 8 Byte messages with 64 processors. No machine achieved binary tree parallelism on this test. There was insufficient memory on the T3E-900 to run this test for a 1 MB message with 64 processors.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	501	164	409	3.1, 1.2
8	4	473	131	317	3.6, 1.5
8	8	460	140	252	3.3, 1.8
8	16	446	169	195	2.6, 2.3
8	32	403	198	124	2.0, 3.3
8	64	357	NA	11	---, 34.
1,000	2	45347	12353	21256	3.7, 2.1
1,000	4	37308	8715	16422	4.3, 2.3
1,000	8	32114	7803	14525	4.2, 2.2
1,000	16	30308	7845	12885	3.9, 2.4
1,000	32	25538	7595	9870	3.4, 2.6
1,000	64	22691	NA	7203	---, 3.2
100,000	2	124055	71446	88417	1.7, 1.4
100,000	4	95157	53724	62195	1.8, 1.5
100,000	8	80456	44058	46044	1.8, 1.7
100,000	16	66720	38426	38629	1.7, 1.7
100,000	32	57970	33269	29630	1.7, 2.0
100,000	64	53099	NA	21882	---, 2.4
1,000,000	2	124414	85040	89249	1.5, 1.4
1,000,000	4	94694	60731	52238	1.6, 1.8
1,000,000	8	78962	49852	35427	1.6, 2.2
1,000,000	16	67429	42806	27668	1.6, 2.4
1,000,000	32	47070	38750	19611	1.2, 2.4
1,000,000	64	NA	NA	12810	

**Table 9:** Log normalized data rates in KB/second for mpi\_bcast followed by mpi\_gather.

**Communication Test 10** (scatter-gather, see table 10):

This test uses mpi\_scatter followed by mpi\_gather and measures communication rates for scattering a message from a root processor and then gathering these messages back to the root processor. The communication rate is calculated by  $2*[8*n*(p-1)]/(\text{wall-clock time})$  and then divided by  $\log(p)$  to obtain the log normalized data rate. The largest size message used for this test is 1 MB because of the large memory requirements of this test when 64 processors are used. There was insufficient memory on the T3E-900 to run this test for a 1 MB message with 64 processors.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	386	157	408	2.5, 0.9
8	4	363	132	242	2.8, 1.5
8	8	338	147	187	2.3, 1.8
8	16	296	176	154	1.7, 1.9
8	32	254	211	118	1.2, 2.2
8	64	210	NA	84	---, 2.5
1,000	2	30926	11848	20367	2.6, 1.5
1,000	4	25679	8232	14504	3.1, 1.8
1,000	8	21945	6421	10584	3.4, 2.1
1,000	16	19860	5906	8063	3.4, 2.5
1,000	32	15928	5165	6076	3.1, 2.6
1,000	64	13619	NA	4872	---, 2.8
100,000	2	89496	62385	71765	1.4, 1.2
100,000	4	72857	44637	48251	1.6, 1.5
100,000	8	59724	31484	32303	1.9, 1.8
100,000	16	47085	23891	24593	2.0, 1.9
100,000	32	39389	19139	16461	2.1, 2.4
100,000	64	33957	NA	10952	---, 3.1
1,000,000	2	90755	74551	65198	1.2, 1.4
1,000,000	4	73833	47996	35057	1.5, 2.1
1,000,000	8	59449	34349	25863	1.7, 2.3
1,000,000	16	48263	25905	18488	1.9, 2.6
1,000,000	32	34460	22047	10931	3.2, 3.2
1,000,000	64	NA	NA	9030	

**Table 10:** Log normalized data rates in KB/second for mpi\_gather followed by mpi\_scatter.

**Communication Test 11** (reduce-scatter, see table 11):

The mpi\_reduce\_scatter routine with the mpi\_sum option is functionally equivalent to first reducing messages on all processors to a root processor and then scattering this reduced message to all processors. This MPI routine could be implemented by a reduce followed by a scatter. However, our communication rate is based on achieving minimal data movement and is calculated by  $8 \cdot n \cdot (p-1) / (\text{wall-clock time})$  and then divided by  $\log(p)$  to obtain the log normalized data rate.

From table 11, notice that the T3E-900 achieves better than binary tree parallelism for all the message sizes tested. The P2SC and Origin achieve better than binary tree parallelism for 100KB and 10 MB messages. Notice that the performance of the T3E-900 significantly drops for messages of size 100 KB and 10 MB relative to the P2SC as also occurred for mpi\_reduce.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IB, T3E/Origin
8	2	249	126	174	2.0, 1.4
8	4	224	65	129	3.5, 1.7
8	8	245	63	121	3.9, 2.0
8	16	266	60	113	4.4, 2.4
8	32	291	56	99	5.2, 2.9
8	64	305	NA	84	---, 3.6
1,000	2	16831	11804	11315	1.4, 1.5
1,000	4	14435	9578	8640	1.5, 1.7
1,000	8	15645	5490	9056	2.8, 1.7
1,000	16	15289	5505	10320	2.8, 1.5
1,000	32	17713	5506	10056	3.2, 2.2
1,000	64	19488	NA	9650	---, 2.0
100,000	2	44596	53894	43496	.82, 1.0
100,000	4	38517	58010	35007	.66, 1.1
100,000	8	42933	76120	40586	.56, 1.1
100,000	16	51289	104648	45773	.49, 1.1
100,000	32	66619	102232	53748	.65, 1.2
100,000	64	91865	NA	55125	---, 1.7
10,000,000	2	45522	62931	28573	.72, 1.6
10,000,000	4	38943	73862	23931	.52, 1.6
10,000,000	8	41582	93413	22881	.45, 1.8
10,000,000	16	51248	148883	27664	.34, 1.9
10,000,000	32	69719	186434	35179	.37, 2.0
10,000,000	64	93996	NA	31574	---, 3.0

**Table 11:** Log normalized data rates in KB/second for mpi\_reduce\_scatter .

The next two communication tests are designed to measure communication between “neighboring” processors for a ring of processors using `mpi_cart_create` (with `reorder = .true.`), `mpi_cart_shift`, and `mpi_sendrecv`.

**Communication Test 12** (right shift, see table 12):

This communication test sends a message from processor  $i$  to processor  $(i+1) \bmod p$ , for  $i = 0, 1, \dots, p-1$ . Observe that the data rates for this test will increase proportionally with  $p$  in an ideal parallel machine. Thus, for communication tests 12 and 13, we define the **normalized data rate** to be  $(\text{total data rate})/p$ . In an ideal parallel computer, the normalized data rate for the above communication would be constant since all communication would be done concurrently. For this test the total data rate is calculated by  $8 \cdot n \cdot p / (\text{wall-clock time})$ .

Table 12 gives the normalized data rates for the above communication in KB/second. Notice that both the T3E-900 and P2SC scale well as the number of processors increase (although there is only data for the P2SC up to 32 processors) since the normalized data rates are roughly constant as the number of processors increases. Observe that table 12 shows normalized data rates and not log normalized data rates and hence exhibiting the high degree of parallelism achieved on this test for all three machines, especially the T3E-900. Notice that the performance of the Origin relative to the T3E-900 becomes much worse as the message size increases. Also observe that the T3E-900 is significantly faster than both of the other machines.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	511	153	208	3.3, 2.5
8	4	495	151	218	3.3, 2.3
8	8	456	145	186	3.1, 2.5
8	16	390	139	115	2.8, 3.4
8	32	371	127	62	2.9, 6.0
8	64	367	NA	44	---, 8.3
1,000	2	30180	11595	13233	2.6, 2.3
1,000	4	28671	11194	12726	2.6, 2.3
1,000	8	28060	10857	11271	2.6, 2.5
1,000	16	25282	10388	10302	2.4, 2.5
1,000	32	23587	9196	7330	2.6, 3.2
1,000	64	22649	NA	5444	---, 4.2
100,000	2	134075	43345	38899	3.1, 3.4
100,000	4	109326	43796	37696	2.5, 2.9
100,000	8	129071	43076	33903	3.0, 3.8
100,000	16	126089	41126	28046	3.1, 4.5
100,000	32	121208	29113	15779	4.2, 7.7
100,000	64	109857	NA	8133	---, 14.
10,000,000	2	136591	56884	39041	2.4, 3.5
10,000,000	4	106170	54528	38197	1.9, 2.8
10,000,000	8	137824	54381	28481	2.5, 4.8
10,000,000	16	137434	54161	24256	2.5, 5.7
10,000,000	32	136962	47406	13785	2.5, 5.7
10,000,000	64	133701	NA	2683	---, 50.

**Table 12:** Normalized data rates for right shift in KB/second.

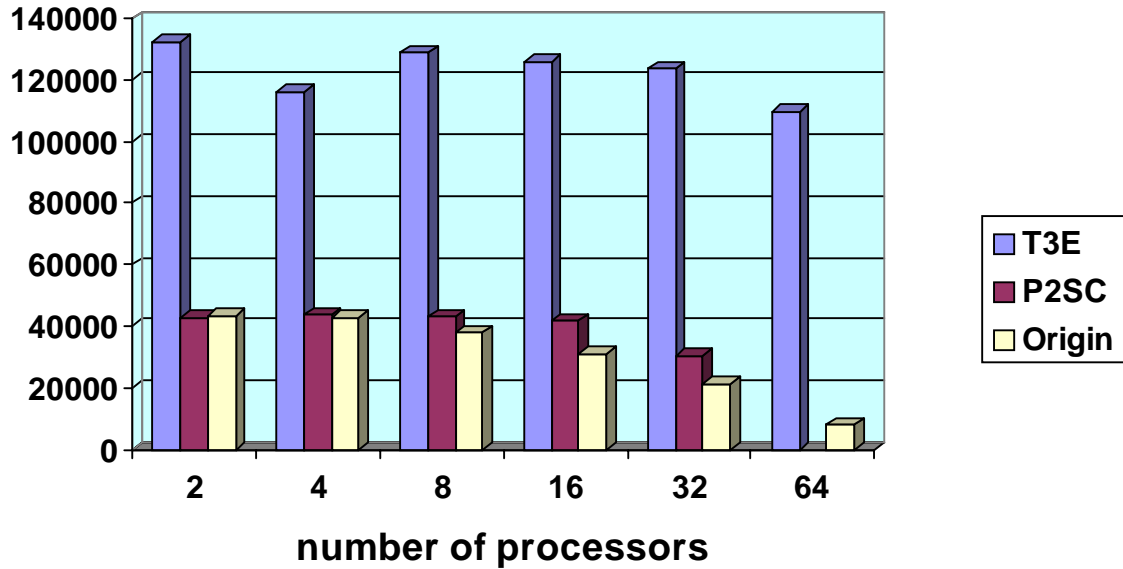
**Communication Test 13** (left & right shift, see table 13 and figure 6):

This test is the same as the above test except here a message is sent from a processor  $i$  to each of its neighbors  $(i-1) \bmod p$  and  $(i+1) \bmod p$ , for  $i = 0, 1, \dots, p$ . Thus, the amount of data being moved on the network will be twice that of the previous test so that the normalized data rate is calculated by  $2 \cdot 8 \cdot n \cdot p / (\text{wall-clock time})$ . Notice that the normalized data rates for communication test 13 are about the same as those for communication test 12. The Origin communication network allows for the concurrent sending of incoming and outgoing streams of

data from one node to another. Because of this, one might expect that the normalized data rates for the Origin for this test to be twice those of the previous test. However, this doubling of the data rate did not occur. Figure 6 shows the normalized data rate for this test for a message of size 100 KB.

Message Size (Bytes)	Number of Processors	T3E-900	IBM P2SC	Origin 2000	T3E/IBM, T3E/Origin
8	2	501	154	202	3.3, 2.5
8	4	484	146	226	3.3, 2.1
8	8	480	143	194	3.4, 2.5
8	16	419	138	117	3.0, 3.6
8	32	393	126	62	3.1, 6.3
8	64	370	NA	44	---, 8.4
1,000	2	29501	11546	13188	2.6, 2.2
1,000	4	27772	10867	12736	2.6, 2.2
1,000	8	28451	10535	11975	2.7, 2.4
1,000	16	26989	10191	10645	2.6, 2.5
1,000	32	23786	8941	7593	2.7, 3.1
1,000	64	23293	NA	5449	---, 4.3
100,000	2	132413	42767	43301	3.1, 3.1
100,000	4	115969	43583	42880	2.7, 2.7
100,000	8	128860	43057	37950	3.0, 3.4
100,000	16	125830	41719	31177	3.0, 4.0
100,000	32	123507	30308	21152	4.1, 5.8
100,000	64	109483	NA	8084	---, 14.
10,000,000	2	142136	54231	29803	2.6, 4.8
10,000,000	4	122336	54339	30085	2.3, 4.1
10,000,000	8	139069	54196	19716	2.6, 7.1
10,000,000	16	139357	53977	16610	2.6, 8.4
10,000,000	32	138746	48236	11938	2.9, 12.
10,000,000	64	123403	NA	2912	---, 42.

**Table 13:** Normalized data rates for the left and right shift in KB/second.



**Figure 6:** Normalized data rates for left and right shifts for a 100 KB message (from table 13).

## CONCLUSIONS

This study was conducted to evaluate relative communication performance of the Cray T3E-900, the Cray Origin 2000 and the IBM P2SC on a collection of 13 communication tests that call MPI routines. Communication tests have been designed to include communication patterns that we feel are likely to occur in scientific programs. Tests were run for messages of size 8 Bytes, 1 KB, 100 KB and 10 MB using 2, 4, 8, 16, 32 and 64 processors (although 64 processors were not available on the P2SC). Because of memory limitations, for some of the tests the 10 MB message size was replaced by messages of size 1 MB. The relative performance of these machines varied depending on the communication test, but overall the T3E-900 was often 2 to 4 times faster than the Origin and P2SC. The Origin and P2SC performed about the same for most of the tests. For a fixed message size the performance of the Origin relative to the T3E-900 would often drop significantly as the number of processors increased. For a fixed message size, the performance of the P2SC relative to the T3E-900 would typically drop as the number of processors increased but this drop was not nearly as much as occurred on the Origin.

## ACKNOWLEDGMENTS

Computer time on the Maui High Performance Computer Center's P2SC was sponsored by the Phillips Laboratory, Air Force Material Command, USAF, under cooperative agreement number F29601-93-2-0001. The views and conclusions contained in this document are those of

the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Phillips Laboratory or the U.S. Government.

We would like to thank Cray Research Inc. for allowing us to use their T3E-900 and Origin 2000 located in Chippewa Falls, Wisconsin and Eagan, Minnesota, USA, respectively.

## REFERENCES

1. *Cray MPP Fortran Reference Manual*, SR 2504 6.2.2, Cray Research, Inc., June 1995.
2. J. Dongarra, R. Whaley, *A User's Guide to the BLACS v1.0*, Computer Science Department Technical Report CS-95-281, University of Tennessee, 1995. (Available as LAPACK Working Note 94 at: <http://www.netlib.org/lapack/lawns/lawn94.ps>)
3. A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, V. Sunderam, *PVM: Parallel Virtual Machine A Users' Guide and Tutorial for Networked Parallel Computing*, The MIT Press, 1994.
4. R. Hockney, M. Berry, *Public International Benchmarks for Parallel Computers: PARKBENCH Committee, Report-1*, February 7, 1994.
5. W. Gropp, E. Lusk, A. Skjellum, *USING MPI*, The MIT Press 1994.
6. G. Luecke, J. Coyle, W. Haque, J. Hoekstra, H. Jespersen, *Performance Comparison of Workstation Clusters for Scientific Computing*, SUPERCOMPUTER, vol XII, no. 2, pp 4-20, March 1996.
7. G. Luecke, J. Coyle, *Comparing the Performance of MPI on the Cray Research T3E and IBM SP-2*, January, 1997 (preprint), see <http://www.public.iastate.edu/~grl/homepage.html>.
8. *Optimization and Tuning Guide for Fortran, C, and C++ for AIX version 4*, second edition, IBM, June 1996.
8. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, J. Dongarra, *MPI: The Complete Reference*, The MIT Press, 1996.
9. <http://www.cray.com>
10. <http://www.austin.ibm.com/hardware/largescale/index.html>